LASR

Leeds Annual Statistical Research Workshop

2006

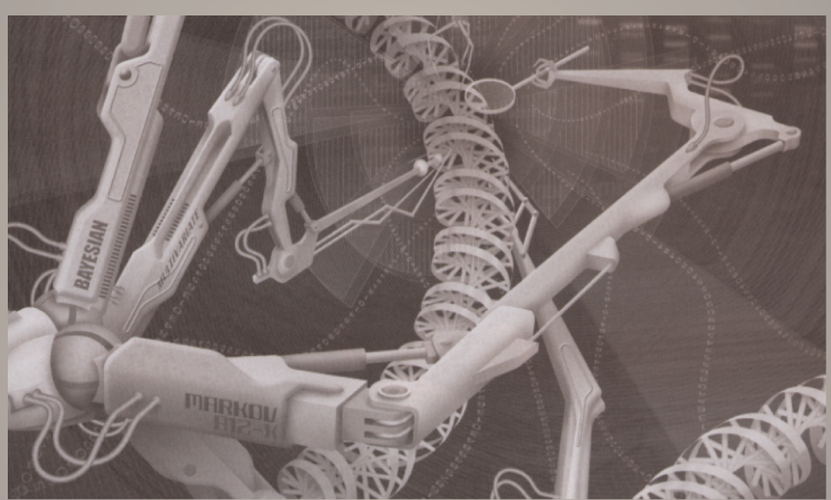

image courtesy of Charles Trevelyan

25th Anniversary

4th - 6th July 2006

interdisciplinary statistics & bioinformatics

Interdisciplinary Statistics and Bioinformatics

International Conference, held in Leeds, UK, 4-6 July 2006,

The 25th Leeds Annual Statistical Research (LASR) Workshop

Sponsored and organized by the Department of Statistics, University of Leeds.

Sponsored by the EPSRC, `http://www.epsrc.ac.uk/`,

and GlaxoSmithKline, `http://www.gsk.com/`

Proceedings edited by

S. Barber, P.D.Baxter, K.V. Mardia, & R.E. Walls

Department of Statistics,

University of Leeds, Leeds LS2 9JT.

Cover design by P. Dickens.

Conference Organizers

S. Barber, P.D. Baxter, R.M. Jackson, J.T. Kent, K.V. Mardia,

J.E. Shuttleworth & D.R. Westhead

ISBN **0 85316 252 2**

Combining geometric morphometrics and genomic data: The genetics of mouse mandible shape

Nicolas Navarro* & Christian P. Klingenberg

Faculty of Life Sciences, University of Manchester

1 Introduction

Newly available genetic designs for quantitative trait locus (QTL) mapping and new genomic resources afford unprecedented statistical power and genetic resolution, and provide new challenges for statistical methods of gene mapping. QTL mapping is the statistical identification (position and effect) of small regions of the genome (ideally, individual genes) that have an effect on a quantitative phenotype.

Since decades, the mouse mandible is used as model for understanding genetic architecture of complex traits. Shape is expected to be affected by many genes of small effect, and according to the functional groups that probably affect mandible shape, the number of shape loci may be several hundreds. Previous experiments using an intercross of two inbreds lines and Procrustes analysis on mandible landmarks found roughly 30 QTLs affecting shape (Klingenberg *et al.* 2004). However, new experimental designs such as heteregeneous stock (HS) are expected to be more powerful by a factor of 30 than this classical intercross (Mott *et al.* 2000).

In this paper, we attempt to map loci affecting the mandible shape using a HS of mice and a dense map of molecular markers. We review some of the problems occuring with high resolution genome data and their consequences on statistical methods and use the geometric information of shape in order to circumvent some of these difficulties.

2 Reconstructing probabilities of ancestral strains

The HS of mice used in this study is a population derived from a pseudo-random breeding scheme of eight inbred strains over 50 generations. In the final generation, each chromosome is a fine grained mosaic of these eight founders (Mott *et al.* 2000). Here, we used individuals from the Northport stock genotyped at $\sim$15K single nucleotide polymorphisms (SNP) from which 12K were informative and without trouble.

Most SNPs have only two states because they result of the replacement of a single base in the DNA sequence. Therefore these markers are unable to ascribe regions of the genome unambiguously to the eight progenitor strains. Nevertheless, an interval-wide probability of the QTL alleles can be obtained from an multipoint dynamic programming algorithm using the HAPPY R package (Mott *et al.* 2000). This interval-wide probability $F_{Li}(s,t)$ is the probability that the individual i descended from the founder strains s and t at marker interval L (Mott *et al.* 2000).

3 Mandible shape mapping

The mandible shape was described by 15 landmarks in 2 dimensions. As far as possible left and right mandibles were digitised, yielding 2,079 individuals with one or both mandibles, of which 1,697 individuals were genotyped. A full generalized Procrutes analysis including a reflection step to take into account the matching symmetry was performed (Dryden and Mardia, 1998; Klingenberg and McIntyre, 1998). The tangent coordinates were averaged over the two sides in order to leave out variation due to asymmetry.

One of the usual mapping methods is to estimate the genetic effects by linear regression of the phenotype on the genotypic scores which are a transformation of the genotype probabilities according to a genetic model (Knott and Haley 2000). This method is especially suitable for dense genomic data because of its computational efficiency. However, this approach is hampered by genetic linkage and sharing of haplotype blocks between founder strains, which induce ill-conditioned genetic matrices and massive collinearity problems. To try to overcome this problem, we used Multivariate Gaussian Mixture Models (MGMM) using an Expectation-Conditional Maximization algorithm (ECM; Jiang and Zeng, 1995). The following animal model is used to estimate the genetic effects of loci:

$$\mathbf{Y} = \boldsymbol{\mu} + \sum_{s,t} \mathbf{F}_L(s,t) \cdot \Phi(\boldsymbol{\beta}_{s,t}, \boldsymbol{\Psi}) + \mathbf{ZU} + \mathbf{E} \tag{1}$$

where $\boldsymbol{\mu}$ is the mean shape, Φ is a multivariate normal distribution with strain or strain combination mean $\boldsymbol{\beta}_{s,t}$ and common covariance matrix $\boldsymbol{\Psi}$, $\mathbf{Z}$ and $\mathbf{U}$ are respectively the design matrix and the effects of covariates (for instance centroid size and gender) and $\mathbf{E}$ is the error effect. According to the genotype class,

$$\boldsymbol{\beta}_{s,t} = \begin{cases} 2\,\mathbf{a}_s & \text{if } s = t \text{ (homozygous genotype)} \\ \mathbf{a}_s + \mathbf{a}_t + \mathbf{d}_{s,t} & \text{if } s \neq t \text{ (heterozygous genotype)} \end{cases} \tag{2}$$

where $\mathbf{a}$ is the additive effect and $\mathbf{d}$ is the dominance effect. Because they are only $q-1$ independent effects over the q alleles, the last additive effect is constrained to $\mathbf{a}_q = -\sum_{s=1}^{q-1} \mathbf{a}_s$. The log likelihood over the n individuals is calculated as

$$\ell(\boldsymbol{\beta}, \boldsymbol{\Psi} | \mathbf{Y}, \mathbf{Z}, \mathbf{M}) = -\frac{n}{2}(2k-4)\log(2\pi) - \frac{n}{2}|\boldsymbol{\Psi}|$$

$$\sum_{i=1}^{n} \log\left(\sum_{s,t} F_{Li}(s,t) \cdot \exp\left(-\frac{1}{2}(\mathbf{Y}_i - \boldsymbol{\mu} - \mathbf{Z}_i\mathbf{U} - \boldsymbol{\beta}_{s,t})^{\mathrm{T}} \boldsymbol{\Psi}^{-} (\mathbf{Y}_i - \boldsymbol{\mu} - \mathbf{Z}_i\mathbf{U} - \boldsymbol{\beta}_{s,t}) \right) \right) \tag{3}$$

where $\mathbf{M}$ stands for all the genetic information use to calculate the $\mathbf{F}_L(s,t)$ probabilities by the multipoint algorithm HAPPY. $\boldsymbol{\Psi}^{-}$ is the Moore-Penrose generalized inverse of $\boldsymbol{\Psi}$ and $|\boldsymbol{\Psi}|$ is obtained as the product of the non null eigenvalues. The presence of a linked QTL is tested using the log likelihood ratio test:

$$-2\log\lambda = 2\Big(\ell(\boldsymbol{\beta}, \boldsymbol{\Psi} | \mathbf{Y}, \mathbf{Z}, \mathbf{M}) - \ell_0(\boldsymbol{\beta} = 0, \boldsymbol{\Psi}_0 | \mathbf{Y}, \mathbf{Z})\Big) \sim \chi^2_{(2k-4)(q+\frac{q(q-1)}{2}-1)} \tag{4}$$

where $\boldsymbol{\Psi}_0 = (\mathbf{Y} - \boldsymbol{\mu} - \mathbf{ZU}_0)^{\mathrm{T}}(\mathbf{Y} - \boldsymbol{\mu} - \mathbf{ZU}_0)$. Following convention, the probabilities were returned as their negative $\log_{10}$ (called thereafter LogP). Their merging yields to a curve of relative significance of linkage along the genome (Figure 1A).

4 Peak selection using LD and geometric information

Because of the computational effort required, the usual approach for controlling the false positives using permutation and tests against the null hypothesis of no effect (Churchill and Doerge, 1994) is difficult if not impossible in the case of MGMM on our dataset. Such LogP threshold, using multivariate regression and only 200 permutations, is equal to 7.3 (1%). Because this threshold is exceeded in most marker intervals, most intervals are associated with an effect and therefore the null hypothesis of no linked effect is rejected in nearly all instances.

An alternative approach for peak selection is to define a marker interval as a peak if this interval is the local maximum of the LogP curve in a region over which there is linkage disequilibrium (LD) on both sides. The LD between marker pairs in the HS drops to average values ($R^2 < 0.5$) within 4 megabases (Mb), and to negligible values ($R^2 < 0.2$) within 8 Mb (Valdar *et al.* submitted). Applied to the scan of the genome, 283 marker intervals (from which 130 are only detected with 4 Mb windows) are retained as local peaks (Figure 1A).

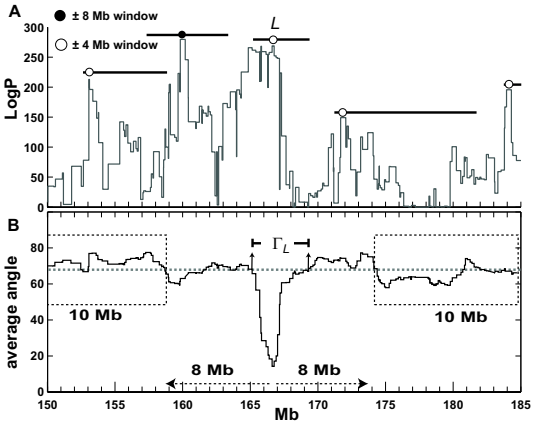

Figure 1: Shape mapping on 35 Mb region of the chromosome 1 using MGMM for the full (additive + dominance) genetic model. The x-axis is the position in megabase pairs (Mb). A. The LogP curve. White and black dots are the selected peaks. Horizontal lines are LD effect intervals on the allelic effects. B. Smoothed average angle of the allelic effects between L and other marker intervals. The grey dotted line is the local asymptote.

Usual confidence intervals on QTL location are derived from 1.5-LogP support intervals, bootstrapping or Bayes credible intervals (e.g., Manichaikul *et al.* submitted). In geometric morphometrics, we can used the information about the direction of the allelic effects in the tangent space. These allelic effects are expected to vary according to the LD. It means that two linked locations tend to have more similar effects than unlinked locations. Therefore, the angle between the allelic vectors is expected to vary from $0°$ to a random angle very near $90°$ in relation to the amount of LD between the compared locations. We proposed to derive for the selected peaks, intervals representing the extent of LD effect on the allelic effects. To do this, we use a relatively simple approach. Firstly, we calculated the average angle $\overline{\alpha}$ between the peak and all other intervals over the q allele vectors in the tangent space:

$$\overline{\alpha} = \sum_{s=1}^{q} \frac{180}{\pi} \cdot \arccos\left(\frac{\mathbf{a}_s^{\lambda}}{||\mathbf{a}_s^{\lambda}||} \cdot \frac{\mathbf{a}_s^{L}}{||\mathbf{a}_s^{L}||}^{\mathrm{T}} \right) / q \tag{5}$$

where λ stands for the varying marker intervals along the chromosome, L stands for the selected peak, $\mathbf{a}_s$ is the additive effect of allele s, and $||.||$ is the Euclidean norm. Then, we smoothed

the $\overline{\alpha}$ curve using an averaging over two marker intervals on the right and on the left plus the location λ. We derived a local asymptote of this curve as the median of the 10 Mb left and right regions 8 Mb apart of the selected peak L (Figure 1B). The extent of the LD on the allelic effects of peak L is then reported as the first left and right crossings of the smoothed curve with the local asymptote. This approach returns intervals in average of 8 Mb ($\pm$5.3).

5 Conclusion

The use of newly available experimental designs and genomic resources such as dense SNP map provides new opportunities and challenges for quantitative trait modelling as well as computational and statistical methodologies. The high resolution mapping of mandible shape loci is possible and yields several hundred potential loci.

Acknowledgments

This work is funded by the Wellcome Trust. We would like to thank Jonathan Flint and Richard Mott for kindly provide the HS mice and their genotypes, and Kirsty Steward and Simon Harold for the preparation of the mandible.

References

Churchill, G. and Doerge, R. (1994). Empirical threshold values for quantitative trait mapping. *Genetics*, **138**, 963-971.

Dryden, I.L. and Mardia, K.V. (1998). *Statistical Shape Analysis*. Chichester, Wiley.

Jiang, C. and Zeng, Z.B. (1995). Multiple Trait Analysis of Genetic Mapping for Quantitative Trait Loci. *Genetics*, **140**, 1111-1127.

Klingenberg, C.P., Leamy, L.J. and Cheverud, J.M. (2004). Integration and modularity of quantitative trait locus effects on geometric shape in the mouse mandible, *Genetics*, **166**, 1909-1921.

Klingenberg, C.P. and McIntyre, G.S. (1998). Geometric morphometrics of developmental instability: Analysing patterns of fluctuating asymmetry with procrustes methods. *Evolution*, **52**, 1363-1375.

Knott, S.A. and Haley, C.S. (2000). Multitrait least squares for quantitative trait loci detection. *Genetics*, **156**, 899-911.

Manichaikul, A., Dupuis, J., Sen, S. and Broman, K.W. (submitted). Poor performance of bootstrap confidence intervals for the location of a quantitative trait locus.

Mott, R., Talbot, C.J., Turri, M.G., Collins, A.C. and Flint, J. (2000). A method for fine mapping quantitative trait loci in outbred animals stocks. *Proceedings of the National Academy of Sciences USA*, **97**, 12649-12654.

Valdar, W., Solberg, L.C., Gauguier, D., Burnett, S., Klenerman, P., Cookson, W.O., Taylor, M., Rawlins, J.N.P., Mott, R. and Flint, J. (submitted). Genome-wide genetic association of complex traits in outbred mice.